4208–4217 *Nucleic Acids Research, 1999, Vol. 27, No. 21*

ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure

Rupert Lück, Stefan Gräf and Gerhard Steger*

Institut für Physikalische Biologie, Geb. 26.12.U1, Heinrich-Heine-Universität Düsseldorf, Universitätsstraße 1, D-40225 Düsseldorf, Germany

Received July 7, 1999; Revised and Accepted September 14, 1999

ABSTRACT

A tool for prediction of conserved secondary structure of a set of homologous single-stranded RNAs is presented. For each RNA of the set the structure distribution is calculated and stored in a base pair probability matrix. Gaps, resulting from a multiple sequence alignment of the RNA set, are introduced into the individual probability matrices. These 'aligned' probability matrices are summed up to give a consensus probability matrix emphasizing the conserved structural elements of the RNA set. Because the multiple sequence alignment is independent of any structural constraints, such an alignment may result in introduction of gaps into the homologous probability matrices that disrupt a common consensus structure. By use of its graphical user interface the presented tool allows the removal of such misalignments, which are easily recognized, from the individual probability matrices by optimizing the sequence alignment with respect to a structural alignment. From the consensus probability matrix a consensus structure is extracted, which is viewable in three different graphical representations. The functionality of the tool is demonstrated using a small set of U7 RNAs, which are involved in 3′-end processing of histone mRNA precursors. Supplementary Material lists further results obtained. Advantages and drawbacks of the tool are discussed in comparison to several other algorithms.

INTRODUCTION

Identification of an RNA structure is a quite demanding task taking into account the enormous number of possible secondary structures, which is about 2^N for a sequence length of N nucleotides (1). To gain insight into the structure–function relationships of single-stranded RNAs despite the complexity of that problem, several experimental (for example enzymatic and chemical mapping, optical melting curves, temperature gradient gel electrophoresis, calorimetry, X-ray studies, NMR, etc.) and computational methods (for example energy minimization, helix list algorithms, genetic algorithms, Monte-Carlo simulations, phylogeny, etc.) have been developed (for reviews see 2–6). Most of the biochemical methods have in common the problem that they are able to determine only the state of the nucleotides, either paired/stacked or non-paired, but not the base pairing partner of a paired nucleotide. Most biophysical methods allow one to determine only thermodynamic and/or kinetic parameters describing the structure but give no detailed information. Thus computational methods are necessary to propose significant structural models that might be verified or rejected by the experimental methods.

With a phylogenetic approach, or comparative sequence analysis, RNA structures are established by selecting from a list of all possible helices those helices that are supported by 'consensus' base changes; i.e. a base pair of a helix in one sequence is changed to another base pair in the same helix of a different sequence. The major problem of such an approach is the need for many sequences, because on the one hand the sequences have to be highly homologous for success in the search for a 'same' helix, and on the other they have to be quite divergent to deliver enough base pair changes to reach statistical significance. This might be even worse when 'the structure alignment does not necessarily reflect the evolutionary relationship between the nucleotides' (7); i.e. a 'correct' sequence alignment does not have to coincide with a 'correct' structure alignment, as shown by van Duin *et al.* (7) in their work on coliphages.

In contrast, the thermodynamic approach based on energy minimization (8–15) needs only a single sequence to find the optimal, many suboptimal, or even the partition function, but is hampered by two assumptions: (i) the structure of the RNA is in thermodynamic equilibrium, which is, for example, not true during or directly after synthesis or for long sequences; (ii) all thermodynamic parameters for structure formation are known with a sufficient degree of accuracy.

Taking into account the merits and drawbacks of the phylogenetic and the thermodynamic approaches, a combination of both methods (16–20) should increase the accuracy of each single method and should help to overcome their individual limitations. We have presented such an algorithm (19) that consists of the following steps (see steps I–V in Fig. 1). (I) For each RNA from a set of homologous RNAs a thermodynamic structure distribution is calculated by energy minimization; the distribution is presented in a matrix of base pairing probabilities (similar to a dot plot). (II) A multiple sequence alignment of the RNA set is produced. (III) Into each of the individual base pairing matrices gaps are introduced as proposed by the sequence alignment. (IV) The resulting 'aligned' matrices are

*To whom correspondence should be addressed. Tel: +49 211 81 14927; Fax: +49 211 81 15167; Email: steger@biophys.uni-duesseldorf.de

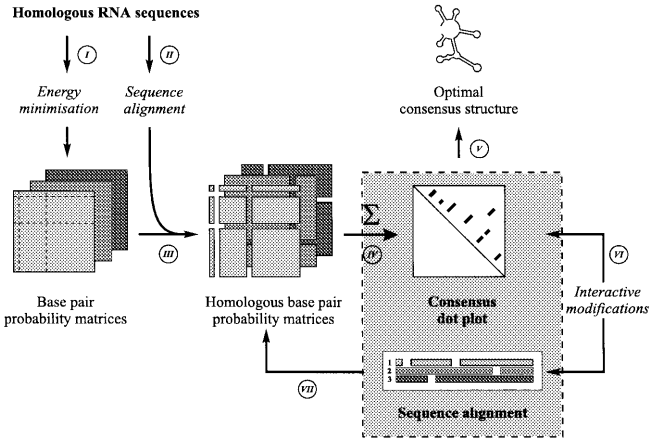

Figure 1. Flow chart of the tool *ConStruct* for determination of a conserved secondary structure. For details see text.

summed up to give a consensus matrix. (V) From the consensus matrix a consensus structure is extracted.

As shown in the publication of the original algorithm (19), it behaves successfully as expected, but there was obviously a problem caused by the multiple sequence alignment (step II) in at least two cases. First, if there is a sequence duplication in a sequence, it is of no importance to a sequence alignment whether the gap (compensating for the missing sequence part in the other sequences) is aligned to the 5′- or the 3′-part of the duplication, but this might be of importance for the consensus structure. Second, special sequences might be completely unrelated, and thus the basis for introduction of gaps, but form identical structural elements; for example, the sequences 5′-GNRA-3′ and 5′-UNCG-3′ (where N is any nucleotide and R is a purine) show no sequence similarity but both are the basis for thermodynamically extra-stable tetraloop hairpins (21,22). Such types of structural misalignments result in non-identical positions of homologous structural elements in the 'aligned' matrices, and are easily recognized by parallel shifts of helices for some of the sequences against similar helices in the remaining sequences.

To allow for the user to resolve such ambiguities and/or to optimize the sequence alignments with respect to a common structure, we have developed a tool including a graphical user interface, which we will describe in the following. In principle this tool allows the user to choose interactively alternative gaps from the pool of suboptimal sequence alignments, which are probably less optimal in terms of sequence alignment, and to trade them with gain against an optimal common structural alignment; any modification of the sequence alignment immediately results in a new representation of the aligned matrices and of the consensus matrix (see steps VI and VII in Fig. 1). The functionality of the tool will be shown for a set of U7 small nucleolar RNAs. This RNA is used as an example because of its short sequence length, which allows us to show screen copies of the full dot plot matrices and not only enlarged portions thereof.

SYSTEMS AND METHODS

RNAfold

csRNAfold is based on *RNAfold* v.1.21 that is part of the Vienna RNA package (13,23,24); it calculates the minimum free energy and the partition function of RNA sequences. Two *ConStruct*-specific options were added: the first allows for writing the base pairing matrices in binary format and the second allows for compression of these matrices by *compress*.

tinoco

tinoco produces simple dot plot matrices containing all possible base pairs of an RNA sequence (25). The 'probability' of each base pair is set to 0.5, which allows for the filtering process in step IV.

File formats

RNAfold (23,24) reads sequences in a special format called Vienna format. To allow for conversion between different sequence formats we added to *readseq* (26) the ability to read and write the Vienna format.

The project file is a simple ASCII file that is created by *cs_make* but might be modified by any text editor. It contains the project name, the file name of the multiple sequence file, and the file names of the base pairing matrices.

The files containing mapping data, depictable by *Circles* (see Fig. 7), have a file name identical to the corresponding sequence but with extension '.map'. The content of these files are multiples of two lines: each first line gives the color of the triangles; each second line gives the nucleotide positions of the triangles separated by blanks. The nucleotide positions have to be those of the unaligned sequence; *Circles* introduces gaps according to the actual alignment. Comment lines (marked by an exclamation mark in the first row) may be interspersed freely.

All relevant windows (i.e. all *tk* canvas widgets) may be printed directly to a *PostScript*-capable printer or saved to a file in *PostScript* format, which should allow for conversion into any graphics format.

Calculation of mutual information content

From the alignment the mutual information content (27–30) of paired positions in the consensus structure is calculated by:

$$I_{x_i,x_j} = \sum_{x_i,x_j} f_{x_i,x_j} \log_b(f_{x_i,x_j}/f_{x_i} f_{x_j})$$

where f_{x_i}, f_{x_j} and f_{x_i,x_j} are the frequencies of the pairing nucleotides x_i and x_j at positions i and j and the joint frequency, respectively. The frequencies of the nucleotides might be corrected for low numbers of sequences or highly conserved positions using the unbiased probability estimator instead of the maximum likelihood estimation (31). The alphabet of allowed nucleotide symbols is A, G, C, U, gap and N; the latter two are taken into account only if present at position i or j. Nucleotides x_i and x_j are statistically interdependent if $I_{x_i,x_j} \geq \chi^2/2M$, where χ^2 is the tabulated χ^2 value with 9, 16 or 25 degrees of freedom for 4, 5 or 6 different alphabet symbols, respectively. For comparison with other publications the basis b of the logarithm may be chosen to be either e or 2.

System

The *ConStruct* package has been tested on a Silicon Graphics Indy and on several different PCs with a Linux operating system (details are listed in Table 1).

The graphical user interface is written using the command language and its corresponding graphics toolkit *tk*/*tcl* (32). *dashpatch* (33) is required to implement an additional canvas

option in the *tk*/*tcl* source necessary for hiding/displaying individual canvas items. The routines for reading of base pair matrices/insertion of gaps in step III and the routine *DrawStructure* are implemented as *C* extensions of the *tcl* interpreter.

Table 1. Hardware and software with which *ConStruct* was tested

Workstation	OS	*tcl*	*tk*	Compiler	*compress*/ *zcat*
SGI Indy	IRIX 6.2	7.5i	4.1i	gcc 2.7.2.2 or cc	4.0
586	Linux 2.0.32	8.0p2	8.0p2	gcc 2.7.2.3	4.2.4

A *C* compiler is required for compilation of *readseq*, of the modified *RNAfold*, and of a few routines that enhance *tk*/*tcl* with the ability to read the base pair probability matrices and to produce the structure representations.

If a compression/decompression program like *compress* or *zcat* is available, *csRNAfold* uses them to compress the base pair probability matrices prior to storage, and the corresponding *cs_dp* routine uses *zcat* to expand the matrices. This is convenient to save disk space.

The complete *ConStruct* package, as described in this paper, is available from http://www.biophys.uni-duesseldorf.de/local/ConStruct.html

Sequences

The *Xenopus borealis* and *Xenopus laevis* sequences were taken from two EMBL entries (34,35); each of these contains a cluster of U7 RNA sequences; the names given in the figures and text are xb#1_#2 and xl#1_#2, respectively, with #1 and #2 the positions of the start and end nucleotide, respectively. For all other sequences the given names are identical to the EMBL IDs (36–40).

RESULTS

Algorithm

Base pair matrices for each RNA of a set of homologous RNAs are calculated either by energy minimization or by using a simple dot plot procedure. The sequences are aligned to identify homologous regions. Then a consensus structure is calculated by extracting structural elements common to all sequences. The algorithm consists of the following steps (see Fig. 1).

Step I. For each sequence R_k with sequence length n_k from a set of homologous RNAs $R_1 \cdots R_M$, a 2-dimensional base pair matrix is created. This matrix is either a simple dot plot showing all possible base pairs (see *tinoco* in Systems and Methods) or a base pair probability plot calculated by one of the energy minimization algorithms, *RNAfold* (13,23) or *LinAll* (11,41), known from the literature. With *RNAfold* the total structure distribution is calculated and stored in a matrix; with *LinAll* the optimal and a definite number of suboptimal structures, which are sufficient to represent the structure distribution, are calculated and stored. In both cases the matrices of base pair probabilities account for thermodynamic weighting of structural alternatives. Each matrix can be viewed as a dot plot. With *RNAfold* or with *LinAll*, the area of a dot is proportional to the base pair probability of the nucleotides at the corresponding position i,j with $1 \leq i < j \leq n_k$. With the simple base pair dot plot each pair in the matrix has the same 'probability'. Execution of the calculations is simplified by the small program *cs_make* (for a further description see below and Fig. 2).

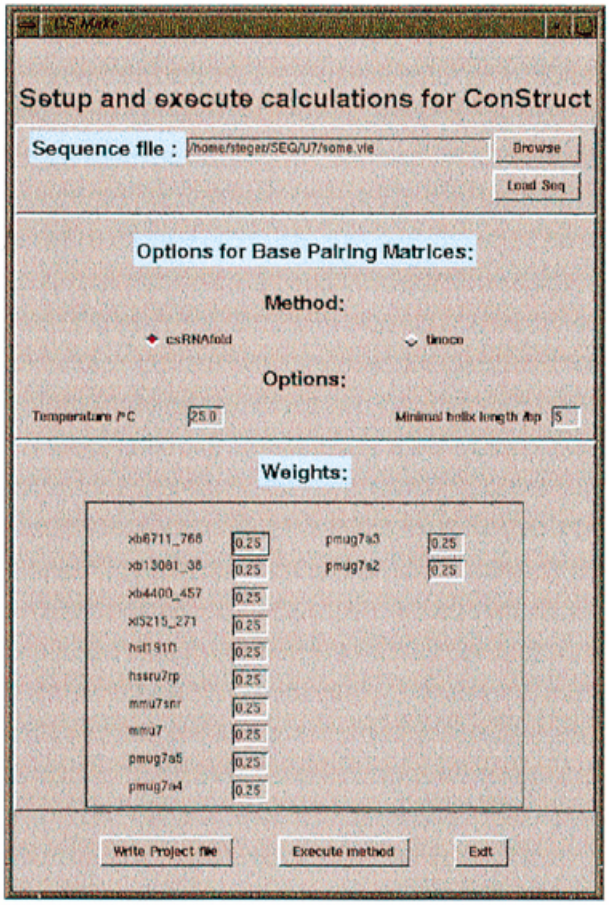

Figure 2. *cs_make*. A graphical user interface allows the user to load a multiple sequence file containing the aligned sequences (top) and to select a program for calculating the base pairing matrices, either by energy minimization with *csRNAfold* or by a dot plot with *tinoco* (middle). After loading the sequence file, a further window pops up (bottom) that allows the user to enter values for weighting the different sequences (see text, step IV). The buttons allow the user to execute the selected program for the loaded sequences in background mode, to write a project file, which is used to load the base pairing matrices into *cs_dp* (see Fig. 3), and to exit *cs_make*. In this example 12 U7 sequences were loaded, matrices were calculated by *csRNAfold* at 25°C, and the project file was stored with equal weights for each of the four RNAs belonging to three groups (x*, *Xenopus*; hs* and mmu7*, mammals; pmug7a*, sea urchin).

Step II. With a cluster alignment program [for example *ClustAl* (42–44) or *PileUp* (45)] a multiple sequence alignment is created for the set of RNAs. The aligned sequences are of identical length *N*, which is larger than their original size n_k due to insertion of gaps. Homologous nucleotides should have the same positions in all aligned sequences.

Step III. Gaps are introduced into the base pair matrices at positions corresponding to the gap positions of the aligned sequences. This results in 'homologous' or 'aligned' base pair matrices of

dimensions $N \times N$ nucleotides. Within each matrix homologous base pairs should be found at exactly the same location. Take into account, however, that only an alignment of the primary structure was performed during step II. This might lead to a misalignment in terms of the secondary structure.

Step IV. The M homologous base pair matrices are summed so that the base pair probability $p_k(i,j)$ of each sequence k contributes to the conserved base pair probability $p_c(i,j)$:

$$p_c(i,j) = \{[\Sigma_{k=1}^{M} w_k \cdot p_k(i,j)^{1/a}]/[\Sigma_{k=1}^{M} w_k]\}^b$$

Probabilities of each sequence are weighted with a specific, user-definable factor w_k to avoid over-representation of a sequence family in comparison to other sequences. For example (see Supplementary Material, Table S5), having a set of 7SL RNAs with nine hop sequences and one from rice it is appropriate to choose $w_{hop} = 1/9$ and $w_{rice} = 1$. The weight values are attached to the sequences or might be modified with help of the program shown in Figure 2 and described below. The exponents $1/a$ and b in the summation were chosen to suppress individual but not conserved probability values in the matrix. Values $a = 3$ and $b = 3$ are used in the example. In the case of a certain helix being present only in a single or a few of the homologous base pair matrices, the appearance of that helix in the consensus base pair matrix is suppressed by the exponentiation. In contrast, a helix common to most if not all of the homologous base pair matrices at the identical position shows up prominently in the consensus base pair matrix. For example, if only one sequence forms a certain base pair with $p_1(I,J) = 1$ and the nine others of the set are not able to form that base pair [$p_{2\ldots10}(I,J) = 0$], values $a = 1$ and $b = 1$ result in $p_c(I,J) = 0.1$ whereas $a = 3$ and $b = 3$ result in $p_c(I,J) = 0.001$, suppressing the noise.

Step V. From the consensus base pair matrix a consensus structure is extracted by means of dynamic programming and backtracking; that procedure maximizes the sum of base pair probabilities. At present only the optimal consensus structure is generated. We will improve that routine to allow the user to also extract suboptimal folds, which might be of importance in the case of RNAs that use conformational switching to implement their biological function.

The consensus structure might be viewed directly in three different graphical representations. (i) The first representation, plotted by *ConStructAlign*, is basically an alignment of the homologous sequences (see for example Fig. 5); the background of the nucleotide characters is colored according to the nucleotide's structural features: loop regions and dangling ends are light green; base paired regions have a reddish color. Nucleotides coinciding with the consensus sequence and forming a base pair have a red background. Nucleotides in pink differ from the consensus sequence but still form a base pair; that base pair is a consensus mutation supporting the predicted consensus helix. In addition to the graphical representation a text output is displayed that describes the structural alignment in numerical form [number of base pairs, number of consensus base pair changes, number of mismatches, consensus base pairing probability and mutual information content (see Systems and Methods) per helical position]. (ii) The second representation, produced by the program *DrawStructure*, is similar to that one might draw by hand (see for example Fig. 6). The backbone distances in loops and in helices are identical. Two drawing modes for loops are available. With the first mode (see Fig. 6A and B) the two halves of each internal loop bridge identical distances; thus the two helices connected by a loop are collinear, which diminishes the chance for overlap of structural regions. With the second mode (see Fig. 6C) loops are drawn as equiangular polygons. Overlap of helical regions may be avoided by user interaction; each helix is selectable by the mouse and might be rotated upon the upstream loop. (iii) The third representation, plotted by *Circles*, is a circular graph with the nucleotides as edges and base pairs connected by arcs (see for example Fig. 7). If chemical or enzymatic mapping data are available the accessibility of nucleotides might be marked by small triangles. Furthermore, the user may store a file describing the consensus structure, useful as input to further drawing programs like *naview* (46; program available as part of the mfold package, 14), *RnaViz* (47), *XRNA* (48) or others.

Steps VI and VII. Steps III–VII are integrated into the tool *cs_dp* (Fig. 3). According to the file names given in the project file (produced by *cs_make*; see step I and Fig. 2), *cs_dp* loads the file with the aligned sequences and the files containing the individual base pair probability matrices. The gaps from the alignment are introduced into the matrices resulting in the homologous base pairing matrices (step III). After summation and weighting (step IV) *cs_dp* shows on the screen the individual dot plots as well as the consensus dot plot (see Figs 3 and 4) in a single frame. In a second frame the alignment of the sequences is shown. The major advantage of the graphical user interface (GUI) of *cs_dp* is that the position of the base pairs from the dot plots is coupled with the position of the corresponding nucleotides in the alignment (step VII). For example, pointing with the mouse to a consensus base pair highlights these base pairs in the alignment with a color from white to red according to the individual base pairing probabilities (see Fig. 3); pointing to a base pair of a selected sequence highlights the corresponding 5′ and 3′ nucleotide in the alignment in blue and red, respectively (see Fig. 4A); pointing to a base paired nucleotide in the alignment changes the color of the corresponding base pair in the dot plot from green or blue to cyan. A selected region of a single sequence, which neighbors a gap, might be moved with the mouse towards the gap, and dot plot and consensus dot plot are updated correspondingly (step VI). A gap might be inserted or removed from a selected sequence by a button press. All these functions of the GUI might be necessary to align structural elements in different sequences that were misaligned by the pure sequence-oriented alignment tool in step II.

EXAMPLE

Eukaryotic histone gene transcripts do not acquire poly(A) tails; instead the histone mRNAs terminate with a hairpin-like stem–loop structure and a short conserved sequence. These signals are recognized by a small nuclear RNA–protein complex (snRNP) containing the U7 RNA, which has a length of ~60 nt. U7 RNA interacts with the pre-mRNA by forming base pairs with the 3′ sequences. This leads to 3′-end processing to yield the histone mRNA (for reviews see 49,50). We have selected U7 RNA to demonstrate the functionality

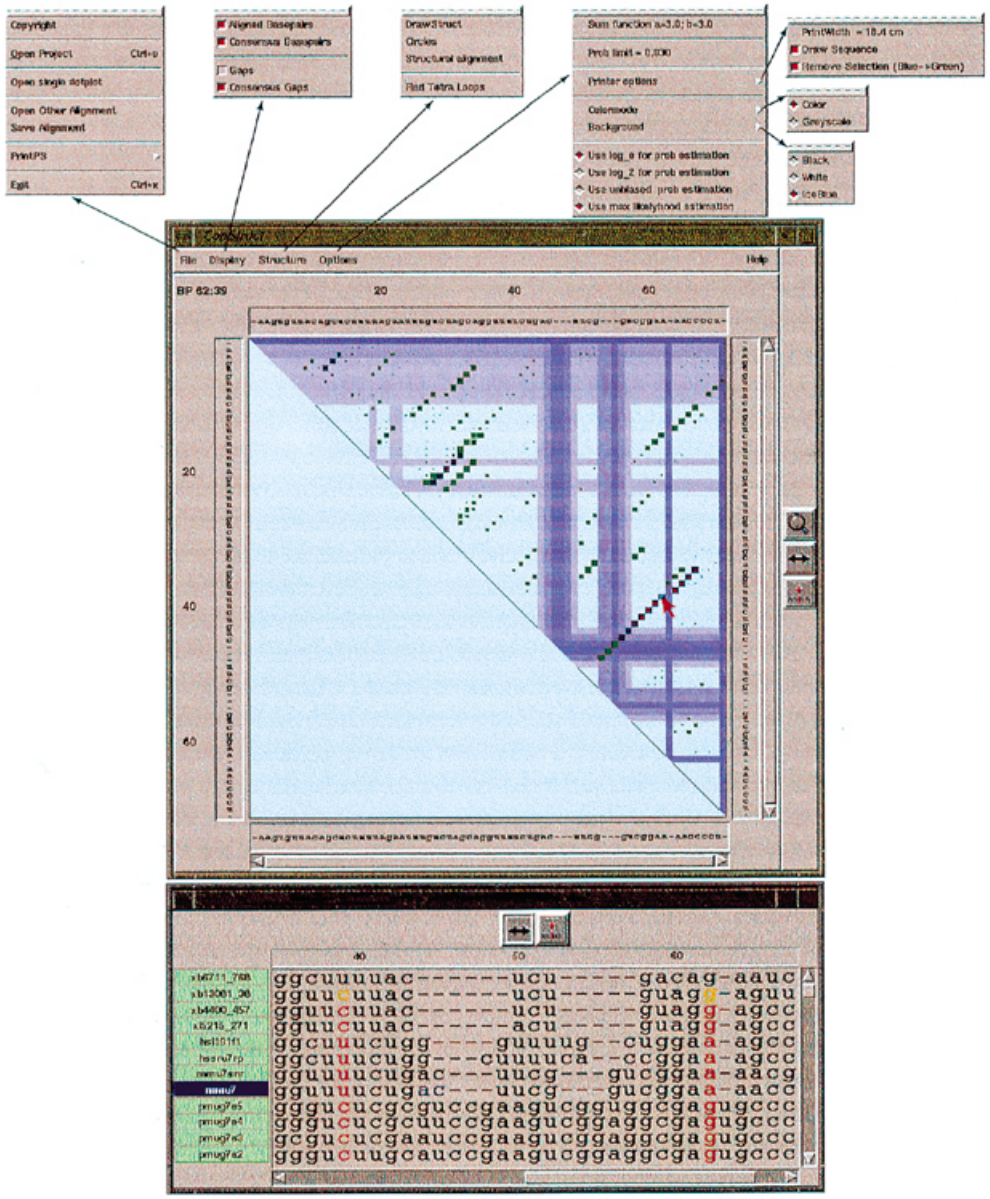

Figure 3. Overlay of base pairing matrices with *cs_dp*. This major tool of *ConStruct* performs steps III–VII of the algorithm (see text and Fig. 1): a project file might be opened (see menu bar, top left), which contains the file name of the aligned sequences, the file names of the corresponding base pair matrices and their weight values (see Fig. 2). During loading of the aligned sequences the gaps from the alignment are introduced into the appropriate matrices (see Fig. 1, step III); this allows for later loading of different alignments. The base pairing probabilities from the individual base pairing matrices are shown as green dots in the dot plot (center); base pairing probabilities of the selected sequence are shown as blue dots. Conserved base pair probabilities (step IV) are shown in white to red color proportional to their probability. Gaps from the alignment are shown as light blue bars; gaps in the selected sequence are shown in white (not done). The buttons in the right border of the dot plot frame and in the top border of the alignment frame allow the user to zoom in/out of the dot plot, to shift to the left or right a selected sequence region from the selected sequence, and to insert/remove a gap. The mouse cursor points to the homologous base pair at position 62:39; this pair is highlighted with reddish color (proportional to the individual base pairing probability) in the alignment (bottom) for all sequences. The sequence mmu7 was selected (note the blue sequence name in the alignment and the blue dots in the dot plot) and its region ^{44}AC^{45}, selected by mouse clicks on the 5′ and 3′ nucleotide, respectively, might be moved in the 3′ direction either for one position by pressing the double-arrowed button with the right mouse button or by moving that region in the alignment directly with the left mouse button up to three positions.

and use of the tool *ConStruct*. The small size of U7 allows full dot plots and alignments to be shown and not only sections thereof; the tool, however, is neither restricted to such short sequences nor to such small sets of RNA as used for that example (see Discussion).

From the EMBL databank (51) at least 26 U7 RNA sequences were available (see also 52). To keep the size of output small and to avoid a bias towards *Xenopus* sequences we selected only 12 sequences from three groups (Amphibia, Mammalia and Echinozoa): three sequences from *Xenopus*

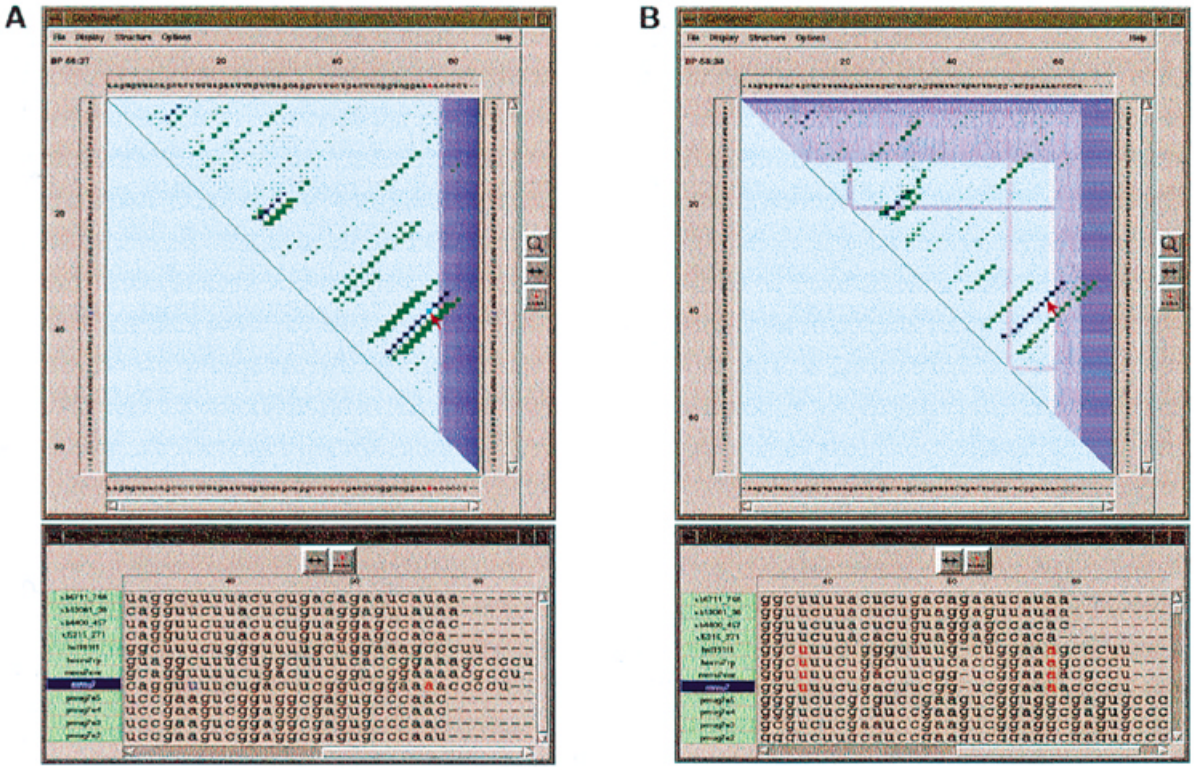

Figure 4. Overlay of base pairing matrices with *cs_dp*. (**A**) The sequences were not aligned; only gaps were added to the 3′-ends to make all sequences equal in length. The mouse cursor points to base pair 56:37 of the sequence mmu7; at that position a base pair exists only in mmu7, as is seen from the highlighting in the alignment (bottom). (**B**) The sequences were aligned by *ClustAl* (43). The mouse cursor points to base pair 58:38 [identical to 56:37 in (A)] of the sequence mmu7; at that position a base pair exists in all mammalian sequences, as is seen from the highlighting in the alignment (bottom).

borealis (34), one from *Xenopus laevis* (35), two from *Homo sapiens* (36,37), two from *Mus musculus* (38,39) and four from *Psammechinus miliaris* (sea urchin) (40). The weights of the individual sequences were set accordingly; base pair matrices were calculated by *RNAfold* at 25°C for Amphibia and Echinozoa and at 37°C for Mammalia (see Fig. 2).

At first the overlay of base pairing matrices was done without an alignment of the sequences; i.e. only the lengths of the sequences were adjusted by gaps at the 3′-end. The result is shown in Figure 4A. As was already obvious from the base pairing matrices of the individual sequences (plots not shown), all sequences, with the exception of xb6711_768, prefer, near their 3′-end, a structure with either a long hairpin or a stem–loop with up to three helices. Because of the missing alignment the structural elements are shifted in parallel between the different sequences. Obviously the structural diversity is much higher at the 5′- than at the 3′-end.

Next, the sequences were aligned with the help of *ClustAlX* (43) using default parameters. The overlay of base pairing matrices, after introduction of the gaps from the alignment, is shown in Figure 4B. As expected the alignment improves the overlay of helices (compare Fig. 4A and B). However, as is obvious from the parallel shift in the different stem–loops near the 3′-end (see position of mouse cursor in Fig. 4B), the alignment is far from optimal in terms of a structural alignment. This might be due to a ‘failure’ of the cluster alignment: the sequences in each group are aligned quite well but the three groups show no alignment. In other words, the similarities in each group are from ~80 to >90%, whereas in between the groups the similarities are only from ~60 down to near 20%.

Lastly, the structural alignment was optimized by hand using the GUI provided by *cs_dp*. In the following we will describe only the optimizations performed for aligning the 3′ stem–loop (compare Fig. 4B with 3). In the *Xenopus* sequences an additional gap was introduced at position 52 and the region from nt 47 up to the 3′-end was shifted by 11 positions downstream. The gap is necessary to compensate for the bulge loop present in the urchin 3′ stem–loop; the shift moves the 3′ sequence of the *Xenopus* helix on top of the urchin 3′ stem–loop. Similarly, in the mammalian sequences a gap was added at position 58 and the region from nt 52 up to the 3′-end was shifted by six positions downstream. In total the alignment was increased in sequence length by one position (the gap mentioned above) and regions between groups, but not in a group, were shifted. Because the alignment between groups had only a marginal basis, the optimization had nearly no effect on the consensus sequence but a dramatic effect on the overlay of structural elements; i.e. the mean base pairing probability of the consensus structure increased from 0.04 to 0.38 per base pair. In the terminal part of the 3′ helix the base pairing probability increased from 0.04 to 0.66; compare the tiny reddish dots in Figure 4B with the large red dots in Figure 3. For a comparison with comparative

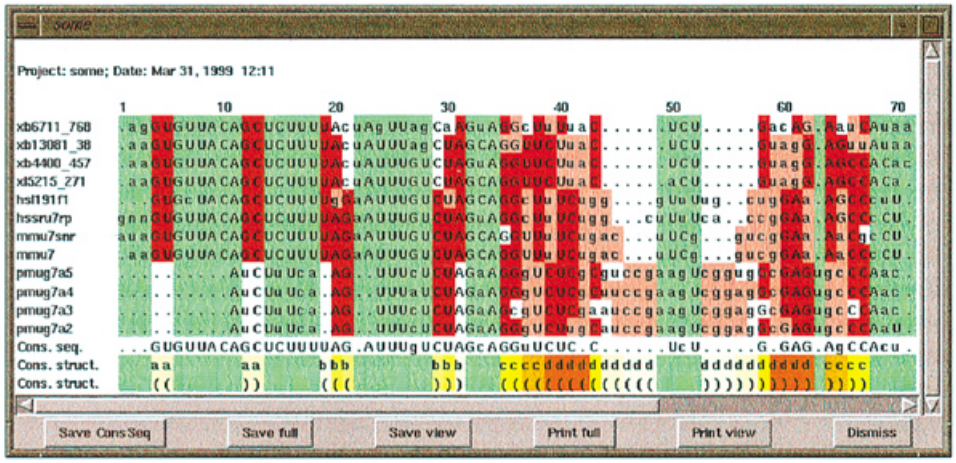

Figure 5. *ConStructAlign*. The alignment of Figure 3 including the consensus structure is shown color coded: regions in light green are non-base paired; regions in white, orange, or light pink are base paired (white 'base pairs' contradict the consensus base pair, light pink base pairs show consensus base pair changes); in the two lines labeled 'Cons. struct.' the consensus structure is shown with increasing base pairing probabilities from white to red. The regions forming a consensus helix are marked either by small characters or with the bracket notation; for example, the regions nt 40–49 and nt 54–63 are base paired with each other and both are marked 'd'.

sequence analysis, the mutual information content of the base pairs in the alignment was checked (27–30). The interdependence of the nucleotides in the helical regions (31) has only a very low significance, mostly due to the low number of sequences. Only after taking all U7 sequences from the database (this sums up to a total of 26 sequences mainly by adding the remaining sequences from the two *Xenopus* clusters; 34,35) the interdependence of the nucleotides in the helical regions reaches χ^2 significance levels of up to 0.99 (see Supplementary Material, Table S1). Both numerical evaluations, base pairing probability and information content, are displayed in a separate, printable text window (not shown).

From the consensus dot plot (see the red dots in Fig. 3) a consensus structure was extracted; a representation of this structure might be shown in three different styles, as given in Figures 5–7.

Figure 5 shows the first representation of the consensus structure in terms of an alignment overlayed by the structural features. According to the coloring scheme of the line marked 'Cons(ensus) struct(ure)' a significant helical region is the proximal part of the 3′ stem–loop: it consists of a 9 bp contiguous helix in *Xenopus*, a 10–11 bp contiguous helix in Mammalia and up to 14 bp interrupted at least by a bulge loop (position 63) in *Psammechinus*. The hairpin loop consists of 3–7 nt; in mouse the loop has the sequence of an extra-stable tetraloop (22).

In the 5′-region the structure, if any consensus structure exists, is much less conserved: in *Xenopus* and Mammalia it consists of two small hairpins; in *Psammechinus* only the second hairpin is possible. There are, however, further possible structural alignments that do not differ significantly in probability from the given alignment. Neither these nor the shown alignment are substantiated by significant numbers of consensus base pair changes.

Figures 6 and 7 show standard representations of the optimal consensus structure with the sequence of mmu7. In both cases the lines connecting base pairs are colored from white to red proportional to the consensus base pairing probability; the color code should help the user to interpret the reliability of individual parts of the consensus structure (53). The first representation is a spider-like graph and the second a circular graph.

DISCUSSION

We have presented here a tool for prediction of conserved secondary structure of a set of homologous RNAs. The tool is based on thermodynamic prediction of the RNA structure distributions but should allow even the inexperienced user to combine the information from thermodynamics with the information from sequence alignment in an intuitive way.

The simple generation and handling of all that information, driven by a GUI, is demonstrated with U7 RNA. The obtained result is in line with the literature (49,50). Briefly, the 3′-region of U7 forms a thermodynamically stable stem–loop structure while there is no common structure in the 5′-region, which has to interact with the histone pre-mRNA.

Despite the example, which shows the use of *ConStruct* for a very short RNA and only a few sequences, *ConStruct* has only a few limitations in that respect: *RNAfold*'s computational and storage effort is $O(n_k^3)$ and $O(n_k^2)$, respectively, for each of the M sequences with individual sequence lengths n_k; with *cs_dp* the memory requirement is $O(M \times N^2)$ and the computational effort to move a subsequence of length n, which is usually much smaller than the alignment length N, is $O(M \times n^2)$. In summary, the requirements allow for easy handling of at least 20 sequences of lengths up to 1000 nt. For example, using *ConStruct*, the consensus structure prediction for PrP mRNAs was done with 23 sequences and an aligned sequence length of ~800 nt (19), for the hepatitis B virus post-transcriptional regulatory element with 30 sequences and a length of ~630 nt (54), for U3 snRNAs with 36 sequences and a length of

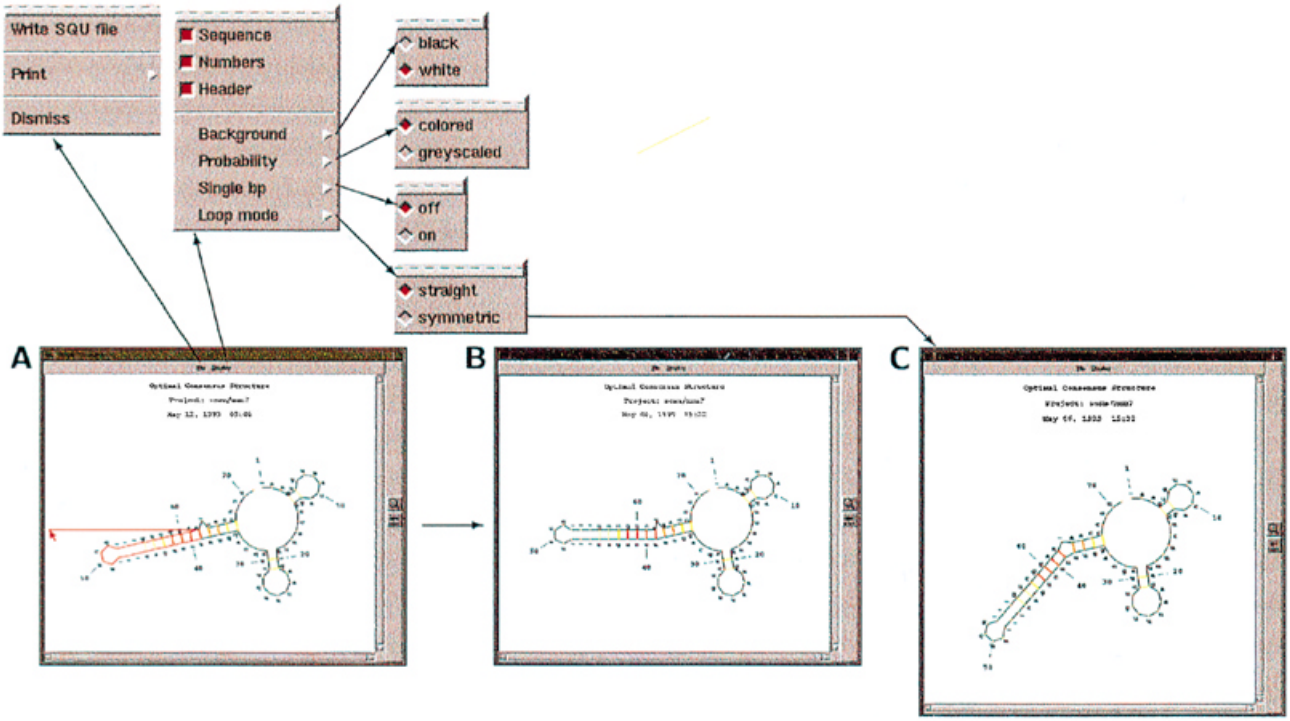

Figure 6. *DrawStructure*. The consensus structure, as extracted from the consensus dot plot of Figure 3, is depicted with colored lines connecting base pairs; the color from white to red shows increasing probability of base pairing. With the mouse the helix 40–49/54–63 was selected [note the red 'backbone' in (A)] and will be bent by ~15° (note the red line pointed to by the mouse). After releasing the mouse button that helix is bent accordingly (B). In (**A**) and (**B**) bulge and internal loops are drawn in such a way that neighboring helices are collinear; in (**C**) loops are drawn as equiangular polygons.

~340 nt, and for plant 7SL RNAs with 18 sequences and a length of ~330 nt (55).

ConStruct needs only a few sequences, much less than are necessary for comparative/phylogenetic structure prediction, to produce a quite convincing consensus structure. Trivially, with energy minimization one needs only a single sequence to get a result, and any further different but homologous sequence adds information to the consensus dot plot and structure. How many sequences are necessary for a certain problem depends on the result of the energy minimization, on the quality of the alignment, and on the diversity of the sequences, but we are not able to give a qualified rule for the sequence number. For the U7 RNAs, for example, 12 or even fewer sequences were sufficient to come up with a consensus structure, but statistical significance was reached with only about twice as many sequences.

On the other hand, any structure prediction algorithm based on thermodynamics will fail when a (sub)structure depends significantly on non-standard base pairings, for which thermodynamic parameters are not known. For example, *ConStruct* predicts for domain IV of plant 7SL RNAs a stem–loop of 12 conventional stacks (55), whereas the model based on phylogeny predicts a stem–loop of 13 stacks including four G·A base pairs (56). For plant sequences, however, both models are not supported statistically (see Supplementary Material, Table S5).

csRNAfold, like basic *RNAfold*, allows the user to restrict the calculated structures; i.e. specific constraints may be used to force the formation of certain base pairs or the pairing of certain bases, or to prohibit the pairing of certain bases. Furthermore, the user may allow for additional pairings like G·A pairs. These features were used neither with the example on U7 RNA nor in any of the examples in the Supplementary Material. This might be useful, however, in the case of excellent mapping data or an already proven consensus structure, as shown in the work of Gaspin and Westhof (57,58).

Hofacker *et al*. (20) use a similar procedure as described here in their program *alidot*; the main difference is their use of only optimal (minimum free energy) structures instead of the base pairing probability matrices. Their approach reduces the CPU and storage demand and allows the handling of sequences in the range ~10 000–20 000 nt, but loses any information from suboptimal conformations. For example, the predicted structure distribution of *Xenopus* sequence xb13081_38 is dominated by a thermodynamically optimal Y-shaped structure with base pairings of the 3′- with the 5′-end (note the green dots in the upper right corner of the matrix in Fig. 3; the single pairing probability is up to 0.85). The first suboptimal structure, however, coincides with and contributes a single pairing probability of up to 0.13 to the consensus structure. Thus taking into account suboptimal structures or the structure distribution should improve the quality and accuracy of the prediction, as already mentioned by Hofacker *et al*. (20) in

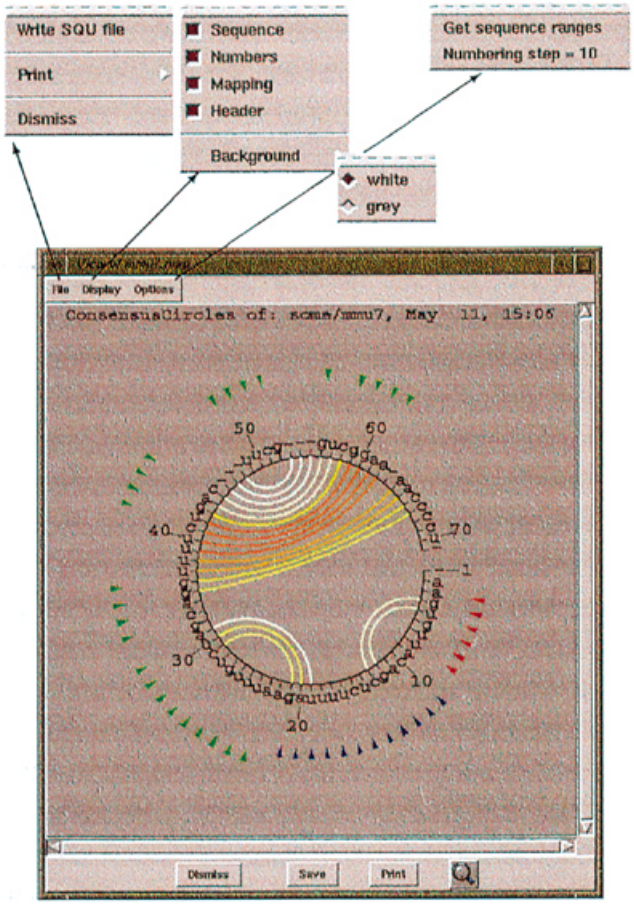

Figure 7. *Circles*. The consensus structure, as extracted from the consensus dot plot of Figure 3, is depicted as a graph with colored arcs connecting base pairs; the color from white to red shows increasing probability of base pairing. The nucleotides of the 3′ stem–loop found to be essential for function of sea urchin U7 RNA are marked by green triangles surrounding the graph; blue triangles mark the region that is recognized sequence-specifically by proteins carrying anti-SM antigenic determinants; red triangles mark the region that is complementary to histone precursor RNA and was shown to be required for function (49). Similarly, mapping data might be read from a file and depicted in the graph.

their discussion. In addition, a thermodynamic structure prediction for sequences of lengths above a few thousand nucleotides which ignores any kinetic influences is not recommended by us, and will only come up with small stable substructures but not with a total structure. Furthermore, *alidot* does not allow for interactive modifications of the sequence alignment, a feature that is usually necessary to obtain the consensus structure. This is shown convincingly with the U7 RNAs (for further examples see Supplementary Material).

ConStruct predicts conserved secondary but not tertiary structure because neither the backtrack in step V nor the basic *RNAfold* are able to handle tertiary interactions. This limitation might be overcome by using an energy minimization algorithm that takes into account pseudoknot or other loop–loop interactions [like those published recently by Stormo's (59) and Eddy's (60) groups]; as a consequence, however, the CPU and storage demand would increase to $M \times N^6$ and $M \times N^4$, respectively, which would make a tool like *ConStruct* non-operable on todays workstations. An alternative would be to use a heuristic algorithm that allows for prediction of tertiary structure, for example those from Pleij's group (61–63).

SUPPLEMENTARY MATERIAL

See Supplementary Material available in NAR Online.

ACKNOWLEDGEMENTS

We thank J. Palinkas for stimulating discussions and Dr D. Riesner for his support. This work was supported by grants from the Deutsche Forschungsgemeinschaft and Fonds der Chemischen Industrie.

REFERENCES

1. Waterman,M.S. (1995) *Introduction to Computational Biology. Maps, Sequences and Genomes*. Chapman & Hall, London, UK.
2. Turner,D.H. and Sugimoto,N. (1988) *Annu. Rev. Biophys. Biophys. Chem.*, **17**, 167–192.
3. Wyatt,J.R., Puglisi,J.D. and Tinoco,I.,Jr (1989) *Bioessays*, **11**, 100–106.
4. Westhof,E. and Michel,F. (1994) In Nagai,K. and Mattaj,I.W. (eds), *RNA–Protein Interactions*. IRL Press, Oxford, UK, pp. 25–51.
5. Schuster,P., Stadler,P.F. and Renner,A. (1997) *Curr. Opin. Struct. Biol.*, **7**, 229–235.
6. Conn,G.L. and Draper,D.E. (1998) *Curr. Opin. Struct. Biol.*, **8**, 278–285.
7. Beekwilder,M.J., Nieuwenhuizen,R. and van Duin,J. (1995) *J. Mol. Biol.*, **247**, 903–917.
8. Nussinov,R., Pieczenik,G., Griggs,J.R. and Kleitman,D.J. (1978) *SIAM J. Appl. Math.*, **35**, 68–82.
9. Waterman,M.S. and Smith,T.F. (1978) *Math. Biosci.*, **42**, 257–266.
10. Zuker,M. and Stiegler,P. (1981) *Nucleic Acids Res.*, **9**, 133–148.
11. Steger,G., Hofmann,H., Förtsch,J., Gross,H.J., Randles,J.W., Sänger,H.L. and Riesner,D. (1984) *J. Biomol. Struct. Dyn.*, **2**, 543–571.
12. Jaeger,J.A., Turner,D.H. and Zuker,M. (1990) *Methods Enzymol.*, **183**, 281–306.
13. McCaskill,J.S.M. (1990) *Biopolymers*, **29**, 1105–1119.
14. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) *J. Mol. Biol.*, **288**, 911–940. http://www.ibc.wustl.edu/~zuker/mfold-3.0.html
15. Zuker,M., Mathews,D.H. and Turner,D.H. (1999) In Barciszewski,J. and Clark,B.F.C. (eds), *RNA Biochemistry and Biotechnology*. NATO ASI Series, Kluwer, Academic Publishers, Dordrecht, The Netherlands. http://www.ibc.wustl.edu/~zuker/seqanal
16. Le,S.Y. and Zuker,M. (1991) *J. Biomol. Struct. Dyn.*, **8**, 1027–1044.
17. Le,S.Y., Zhang,K. and Maizel,J.V.,Jr (1995) *Comput. Biomed. Res.*, **28**, 53–66.
18. Davis,J.P., Janjic,N., Pribnow,D. and Zichi,D.A. (1995) *Nucleic Acids Res.*, **23**, 4471–4479.
19. Lück,R., Steger,G. and Riesner,D. (1996) *J. Mol. Biol.*, **258**, 813–826.
20. Hofacker,I.L., Fekete,M., Flamm,C., Huynen,M.A., Rauscher,S., Stolorz,P.E. and Stadler,P.F. (1998) *Nucleic Acids Res.*, **26**, 3825–3836.
21. Antao,V.P., Lai,S.Y. and Tinoco,I.,Jr (1991) *Nucleic Acids Res.*, **19**, 5901–5905.
22. Antao,V.P. and Tinoco,I.,Jr (1992) *Nucleic Acids Res.*, **20**, 819–824.
23. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,S., Tacker,M. and Schuster,P. (1994) *Monatsh. Chem.*, **125**, 167–188.
24. Hofacker,I. (1998) http://www.tbi.univie.ac.at/~ivo/RNA/
25. Tinoco,I.,Jr, Uhlenbeck,O.C. and Levine,M.D. (1971) *Nature*, **230**, 362–367.
26. Gilbert,D.G. (1993) ftp://ftp.bio.indiana.edu/molbio/readseq/
27. Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) *J. Mol. Biol.*, **188**, 415–431.
28. Gutell,R.R., Power,A., Hertz,G.Z., Putz,E.J. and Stormo,G.D. (1992) *Nucleic Acids Res.*, **20**, 5785–5795.
29. Eddy,S.R. and Durbin,R. (1994) *Nucleic Acids Res.*, **22**, 2079–2088.
30. Gautheret,D., Damberger,S.H. and Gutell,R.R. (1995) *J. Mol. Biol.*, **248**, 27–43.
31. Chiu,D.K. and Kolodziejczak,T. (1991) *Comp. Appl. Biosci.*, **7**, 347–352.

32. Scriptics (1999) http://www.scriptics.com/
33. Nijtmans,J. (1998) http://home.wxs.nl/~nijtmans/
34. Phillips,S.C., Watkins,N.J. and Turner,P.C. (1996) EMBL ID XBU715KB, GenBank accession no. Z54313.
35. Phillips,S.C. and Birnstiel,M.L. (1992) EMBL ID XLU7SNRG, GenBank accession no. X64404.
36. Matthews,P. (1996) EMBL ID HSL191F1, GenBank accession no. Z68756.
37. Mowry,K.L. and Steitz,J.A. (1987) *Science*, **238**, 1682–1687. EMBL ID HSSRU7RP, GenBank accession no. M17910.
38. Cotten,M., Gick,O., Vasserot,A., Schaffner,G. and Birnstiel,M.L. (1988) *EMBO J.*, **7**, 801–808. EMBL ID MMU7SNR, GenBank accession no. X07183.
39. Gruber,A., Soldati,D., Burri,M. and Schümperli,D. (1991) *Biochim. Biophys. Acta*, **1088**, 151–154. EMBL ID MMU7, GenBank accession no. X54748.
40. De Lorenzi,M., Rohrer,U. and Birnstiel,M.L. (1986) *Proc. Natl Acad. Sci. USA*, **83**, 3243–3247. EMBL ID PMUG7A1–PMUG7A5, GenBank accession nos M13307–M13311.
41. Schmitz,M. and Steger,G. (1992) *Comp. Appl. Biosci.*, **8**, 389–399.
42. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.
43. Jeanmougin,F., Thompson,J.D., Gouy,M., Higgins,D.G. and Gibson,T.J. (1998) *Trends Biol. Sci.*, **23**, 403–405.
44. Thompson,J. and Jeanmougin,F. (1998) ftp://ftp.ebi.ac.uk/pub/software/unix/clustalw/
45. Genetics Computer Group (1999) *Wisconsin Package*. Genetics Computer Group, Madison, WI. http://www.gcg.com/
46. Bruccoleri,R.E. and Heinrich,G. (1988) *Comp. Appl. Biosci.*, **4**, 167–173. ftp://snark.wustl.edu/pub/naview.tar.Z
47. De Rijk,P. and De Wachter,R. (1997) *Nucleic Acids Res.*, **25**, 4679–4684. http://rrna.uia.ac.be/rnaviz/
48. Weiser,B. and Noller,H.F. (1999) ftp://fangio.ucsc.edu/pub/XRNA
49. Birnstiel,M.F.L. and Schaufele,F.J. (1988) In Birnstiel,M.L. (ed.), *Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles*. Springer Verlag, Heidelberg, Germany, pp. 155–182.
50. Mowry,K.L. and Steitz,J.A. (1988) *Trends Biol. Sci.*, **13**, 447–451.
51. Stösser,G., Tuli,M.A., Lopez,R. and Sterk,P. (1999) *Nucleic Acids Res.*, **27**, 18–24. http://www.ebi.ac.uk/embl/
52. Zwieb,C. (1998) http://psyche.uthct.edu/dbs/uRNADB/uRNADB.html
53. Zuker,M. and Jacobson,A.B. (1998) *RNA*, **4**, 669–679.
54. Smith,G.J., Donello,J.E., Lück,R., Steger,G. and Hope,T.J. (1998) *Nucleic Acids Res.*, **26**, 4818–4827.
55. Matoušek,J., Junker,V., Vrba,L., Schubert,J., Patzak,J. and Steger,G. (1999) *Gene*, in press.
56. Larsen,N. and Zwieb,C. (1991) *Nucleic Acids Res.*, **19**, 209–215.
57. Gaspin,C. and Westhof,E. (1995) *J. Mol. Biol.*, **254**, 163–174.
58. Chetouani,F., Monestié,P., Thébault,P., Gaspin,C. and Michot,B. (1997) *Nucleic Acids Res.*, **25**, 3514–3522.
59. Tabaska,J.E., Cary,R.B., Gabow,H.N. and Stormo,G.D. (1998) *Bioinformatics*, **14**, 691–699.
60. Rivas,E. and Eddy,S.R. (1999) *J. Mol. Biol.*, **285**, 2053–2068.
61. Abrahams,J.P., van den Berg,M., van Batenburg,E. and Pleij,C. (1990) *Nucleic Acids Res.*, **18**, 3035–3044.
62. van Batenburg,F.H., Gultyaev,A.P. and Pleij,C.W. (1995) *J. Theor. Biol.*, **174**, 269–280.
63. Gultyaev,A.P., van Batenburg,F.H. and Pleij,C.W. (1999) *RNA*, **5**, 609–617.